PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-053395
(43)Date of publication of application : 26.02.1999

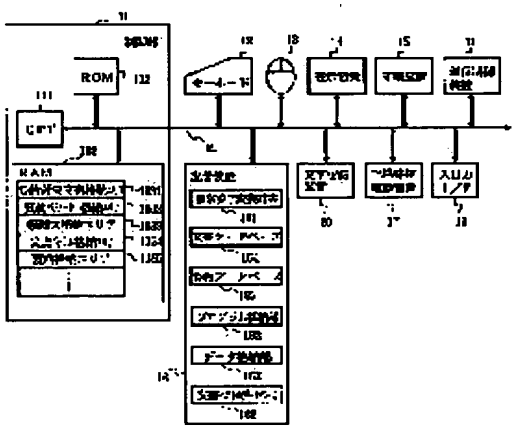
(51)Int.Cl. G06F 17/30

(21)Application number : 09-219300
(22)Date of filing : 29.07.1997

(71)Applicant : JUST SYST CORP
(72)Inventor : NOMURA NAOYUKI
FUJISAWA SHINJI

(54) DEVICE AND METHOD FOR DOCUMENT PROCESSING AND STORAGE MEDIUM STORING DOCUMENT PROCESSING PROGRAM

(57)Abstract:
PROBLEM TO BE SOLVED: To extract the citation relation among plural documents and to produce the summaries including the proper conjunctions and connective sentences by adding the connective words among the summaries produced by a summary production means for the production of summaries of documents.
SOLUTION: A CPU 111 decides the importance of every candidate word (phrase) based on the summary parameters stored in a RAM 16 and the frequency of appearance, evaluation functions, etc., of extracted candidate words (phrases) set in every document group. Then the CPU 111 decides the importance of every sentence included in every document group based on the importance of every candidate word (phrase), the summary parameters, etc., and lists up the sentences within a summary ratio in the order of higher importance decided for the sentences. Furthermore, the CPU 111 arranges the listed-up sentences in the order of appearance set in the document groups to acquire a summary of the relevant document group. When the production of a summary is over to every document, the CPU 111 puts the conjunctions and connective sentences among the A produced summaries to produce a summary sentence.



LEGAL STATUS

[Date of request for examination] 28.07.2004

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平11－53395

(43)公開日　平成11年(1999)２月26日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/401	３２０Ａ
			15/40	３７０Ａ

審査請求　未請求　請求項の数10　ＦＤ　（全９頁）

(21)出願番号　特願平9－219300

(22)出願日　平成９年(1997)７月29日

(71)出願人　390024350
株式会社ジャストシステム
徳島県徳島市沖浜東３－46

(72)発明者　野村　直之
徳島県徳島市沖浜東３丁目46番地　株式会社ジャストシステム内

(72)発明者　藤澤　信二
徳島県徳島市沖浜東３丁目46番地　株式会社ジャストシステム内

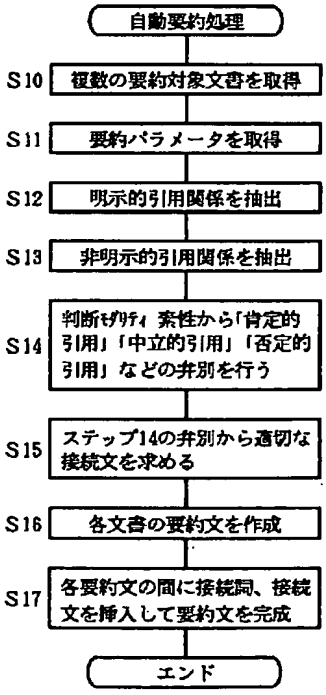
(74)代理人　弁理士　川井　隆　（外１名）

(54)【発明の名称】　文書処理装置、文書処理プログラムが記憶された記憶媒体および文書処理方法

(57)【要約】

【課題】　所与の複数の文書におけるお互いの引用関係を抽出して、適切な接続詞、接続文を含んだ要約を作成する文書処理装置を提供すること。

【解決手段】　複数の文書間の関連性を判定するために、明示的引用関係と非明示的引用関係記述を抽出する。非明示的引用関係記述を抽出するためには、各毎の文書ベクトルを求め、各文書間で文書ベクトルの差をとる。これらの連続する２つの文書間のコサインバリュー（cosine value）が高いか低いかで非明示的引用関係があるか否かを判断する。そして、引用関係記述を肯定的引用、否定的引用、中立的引用に分け、接続文を生成する。そして、各文書を要約し、その文書間に接続文を挿入して、読み易い要約文を生成する。



【特許請求の範囲】
【請求項１】　所定形式の文書を複数個取得する文書取得手段と、
前記文書取得手段により取得された複数の各文書間の関係を認識する関係認識手段と、
前記関係認識手段で認識された関係に対応して各文書間に挿入する接続語を生成する接続語生成手段と、
各文書の要約を自動的に作成する要約作成手段とを備え、
前記要約作成手段により作成された各要約間に前記接続語生成手段により生成された接続語を配置して複数文書の要約を作成することを特徴とする文書処理装置。
【請求項２】　前記関係認識手段により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、
この非明示的引用関係を検知するために各文書の類似性を判断する類似度判定手段をさらに備えたことを特徴とする請求項１記載の文書処理装置。
【請求項３】　前記文書取得手段で取得された複数個の各文書を特徴づける文書ベクトルを決定する文書ベクトル決定手段を備え、
前記類似度判定手段は前記文書ベクトル決定手段で決定された各文書の文書ベクトルにより各文書間の類似度を判定することを特徴とする請求項２記載の文書処理装置。
【請求項４】　前記関係認識手段により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この引用関係を肯定的引用、否定的引用、中立的引用に分類して認識することを特徴とする請求項２または請求項３記載の文書処理装置。
【請求項５】　所定形式の文書を複数個取得する文書取得機能と、
前記文書取得機能により取得された複数の各文書間の関係を認識する関係認識機能と、
前記関係認識出機能で認識された関係に対応して各文書間に挿入する接続語を生成する接続語生成機能と、
各文書の要約を自動的に作成する要約作成機能とを備え、
前記要約作成機能により作成された各要約間に前記接続語生成機能により生成された接続語を配置して複数文書の要約を作成することをコンピュータに実現させるためのコンピュータ読取り可能な文書処理プログラムが記憶された記憶媒体。
【請求項６】　前記関係認識機能により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、
この非明示的引用関係を検知するために各文書の類似性を判断する類似度判定機能をさらに備えたことを特徴とする請求項５記載の文書処理プログラムが記憶された記憶媒体。
【請求項７】　前記文書取得手段で取得された複数個の各文書を特徴づける文書ベクトルを決定する文書ベクトル決定機能を備え、
前記類似度判定機能は前記文書ベクトル決定手段で決定された各文書の文書ベクトルにより各文書間の類似度を判定することを特徴とする請求項６記載の文書処理プログラムが記憶された記憶媒体。
【請求項８】　前記関係認識機能により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この引用関係を肯定的引用、否定的引用、中立的引用に分類して認識することを特徴とする請求項６または請求項７記載の文書処理プログラムが記憶された記憶媒体。
【請求項９】　所定形式の文書を複数個取得し、
取得された複数の各文書間の関係を認識し、
認識された関係に対応して各文書間に挿入する接続語を生成し、
各文書の要約を自動的に作成し、
作成した各要約間に生成した接続語を配置して複数文書の要約を作成することを特徴とする文書処理方法。
【請求項１０】　取得された複数の各文書間の関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、
この非明示的引用関係を検知するために各文書の類似性を判断することを特徴とする請求項９記載の文書処理方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】この発明は、文書処理装置、文書処理方法および文書処理プログラムを記憶した記憶媒体に係り、詳細には、複数の文書から文書間の関係に言及した要約を作成する技術に関する。
【０００２】
【従来の技術】従来、書籍、論文、報告書等の各種の文書に対し、要約（抄録を含む）の自動作成処理や、他文書等との関連づけ処理等の各種処理をコンピュータを用いて行うことが行われている。文書の自動要約については、例えば、「全文情報からの意味的情報の抽出と加工」（情報処理学会第３８回全国大会予稿集、第２２２頁；１９８９年）で提案されている。この方法では、まず文書中の重要語を字種や動詞等の情報から抽出し、さらに重要語の出現頻度から最重要語を決定する。次に、重要語と最重要語が出現するか否かから重要文を決定することで、自動的に要約を作成することが可能になる。また、文章の段落の性質を反映させることで、より正確

に要約を作成する特開平３－１９１４７５号公報に記載された方法等も提案されている。一方、他のデータとの関連づけとしては、インターネットにおけるハイパーリンクや、フレームシステム等による知識処理（エキスパートシステム等）における関連づけ等が行われている。

【０００３】

【発明が解決しようとする課題】ところで、従来の文書処理装置では、単数の文書を対象として、要約するものであった。そこで、複数の文書について要約を作成する場合、個々の文書を要約してこれを繋ぎ合わせることが必要であった。しかし、この方法によると、複数の各文書が同一のトピックのみで構成されている場合は、比較的適切な要約を作成することが可能であるが、各文書のが異なる複数のトピックを含むときは、必ずしも適切な要約を作成することができなかった。すなわち、各文書の内容の類否を考慮せず、互いに異なる主張や事実の記載をもつ複数文書の要約を互いにつなぎ合わせることで要約を作成していたため、可読性の低い要約を生成していた。

【０００４】そこで、本発明は、このような従来の課題を解決するためになされたもので、所与の複数の文書におけるお互いの引用関係を抽出して、適切な接続詞、接続文を含んだ要約を作成する文書処理装置および文書作成方法を提供することを第１の目的とする。また、本発明は、所与の複数の文書におけるお互いの引用関係を抽出して、適切な接続詞、接続文を含んだ要約を作成することができるコンピュータ読取り可能な文書処理プログラムを記憶した記憶媒体を提供することを第２の目的とする。

【０００５】

【課題を解決するための手段】請求項１記載の発明では、文書処理装置が、所定形式の文書を複数個取得する文書取得手段と、前記文書取得手段により取得された複数の各文書間の関係を認識する関係認識手段と、前記関係認識手段で認識された関係に対応して各文書間に挿入する接続語を生成する接続語生成手段と、各文書の要約を自動的に作成する要約作成手段とを備え、前記要約作成手段により作成された各要約間に前記接続語生成手段により生成された接続語を配置して複数文書の要約を作成することにより前記第１の目的を達成する。

【０００６】請求項２に記載した発明では、請求項１に記載した文書処理装置において、前記関係認識手段により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この非明示的引用関係を検知するために各文書の類似性を判断する類似度判定手段をさらに備えたことにより前記第１の目的を達成する。

【０００７】請求項３に記載した発明では、請求項１または請求項２に記載した文書処理装置において、前記文書取得手段で取得された複数個の各文書を特徴づける文書ベクトルを決定する文書ベクトル決定手段を備え、前記類似度判定手段は前記文書ベクトル決定手段で決定された各文書の文書ベクトルにより各文書間の類似度を判定する。

【０００８】請求項４に記載した発明では、請求項２または請求項３記載の文書処理装置において、前記関係認識手段により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この引用関係を肯定的引用、否定的引用、中立的引用に分類して認識する。

【０００９】請求項５に記載した発明では、文書処理プログラムが記憶された記憶媒体が、所定形式の文書を複数個取得する文書取得機能と、前記文書取得機能により取得された複数の各文書間の関係を認識する関係認識機能と、前記関係認識出機能で認識された関係に対応して各文書間に挿入する接続語を生成する接続語生成機能と、各文書の要約を自動的に作成する要約作成機能とを備え、前記要約作成機能により作成された各要約間に前記接続語生成機能により生成された接続語を配置して複数文書の要約を作成することを実現させることにより前記第２の目的を達成する。

【００１０】請求項６に記載した発明では、請求項５記載の文書処理プログラムが記憶された記憶媒体に、前記関係認識機能により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この非明示的引用関係を検知するために各文書の類似性を判断する類似度判定機能を実現させて前記第２の目的を達成する。

【００１１】請求項７に記載した発明では、請求項６記載の文書処理プログラムが記憶された記憶媒体において、前記文書取得手段で取得された複数個の各文書を特徴づける文書ベクトルを決定する文書ベクトル決定機能を備え、前記類似度判定機能は前記文書ベクトル決定手段で決定された各文書の文書ベクトルにより各文書間の類似度を判定させて前記第２の目的を達成する。

【００１２】請求項８に記載した発明では、請求項６または請求項７記載の文書処理プログラムが記憶された記憶媒体において、前記関係認識機能により関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この引用関係を肯定的引用、否定的引用、中立的引用に分類して認識することにより前記第２の目的を達成する。

【００１３】請求項９に記載した発明では、所定形式の文書を複数個取得し、取得された複数の各文書間の関係を認識し、認識された関係に対応して各文書間に挿入する接続語を生成し、各文書の要約を自動的に作成し、作

成した各要約間に生成した接続語を配置して複数文書の要約を作成することにより前記第１の目的を達成する。

【００１４】請求項１０に記載した発明では、請求項９に記載した発明において、取得された複数の各文書間の関係を認識する際、引用関係が明示されている明示的引用関係と、内容の類似性を判定して引用関係が得られる非明示的引用関係を検知し、この非明示的引用関係を検知するために各文書の類似性を判断することにより前記第１の目的を達成する。

【００１５】

【発明の実施の形態】以下、本発明の文書処理装置、文書処理方法および文書処理プログラムを記憶した記憶媒体の好適な実施の形態を、図１ないし図８を参照して詳細に説明する。

（１）実施の形態の概要

本実施の形態では、複数の文書間の関連性を判定するために、明示的引用関係と非明示的引用関係記述を抽出する。非明示的引用関係記述を抽出するためには、各毎の文書ベクトルを求め、各文書間で文書ベクトルの差をとる。これらの連続する２つの文書間のコサインバリュー（cosine value）が高いか低いかで非明示的引用関係があるか否かを判断する。そして、引用関係記述を肯定的引用、否定的引用、中立的引用に分け、接続文を生成する。そして、各文書を要約し、その文書間に接続文を挿入して、読み易い要約文を生成する。

【００１６】（２）実施の形態の詳細

図１は、文書処理装置の構成を表したブロック図である。本実施の形態の文書処理装置は、パーソナルコンピュータやワードプロセッサ等を含むコンピュータシステムとして構成し、また、ＬＡＮ（ローカル・エリア・ネットワーク）のサーバーやインターネットを含むコンピュータ（パソコン）通信のホストとして構成することが可能である。文書処理装置は、図１に示すように装置全体を制御するための制御部１１を備えている。この制御部１１には、データバス等のバスライン２１を介して、入力装置としてのキーボード１２やマウス１３、表示装置１４、印刷装置１５、記憶装置１６、記憶媒体駆動装置１７、通信制御装置１８、および、入出力Ｉ／Ｆ１９、および、文字認識装置２０が接続されている。制御部１１は、ＣＰＵ１１１、ＲＯＭ１１２、ＲＡＭ１１３を備えている。ＲＯＭ１１２は、ＣＰＵ１１１が各種制御や演算を行うための各種プログラムやデータが予め格納されたリードオンリーメモリである。

【００１７】ＲＡＭ１１３は、ＣＰＵ１１１にワーキングメモリとして使用されるランダム・アクセス・メモリである。このＲＡＭ１１３には、本実施の形態による要約処理を行うためのエリアとして、要約対象文書格納エリア１１３１、要約パラメータ格納エリア１１３２、接続文格納エリア１１３３、文書ベクトル格納エリア１１３４、要約格納エリア１１３５、その他の各種エリアが確保されるようになっている。要約パラメータ格納エリア１１３２には、操作者からの入力等により取得された要約パラメータの値または後述のデータ格納部の１６３から読み込んだ要約パラメータのデフォルト値が格納される。操作者が入力する要約パラメータとしては、例えば、全文書に対する要約の比率（１％～９９％）、数量優先のある／なし、長単文のある／なし、です／であるの選択をする／しない、等の値が格納される。接続文格納エリア１１３３には、各要約文を接続するために生成された接続文が格納される。文書ベクトル格納エリア１１３４には、要約対象文書に対する文書ベクトルと、後述する各サブ文書に対する文書ベクトルとが格納される。要約格納エリア１１３５には、各文書の要約文が格納される。

【００１８】キーボード１２は、かな文字を入力するためのかなキーやテンキー、各種機能を実行するための機能キー、カーソルキー、等の各種キーが配置されている。マウス１３は、ポインティングデバイスであり、表示装置１４に表示されたキーやアイコン等を左クリックすることで対応する機能の指定を行う入力装置である。表示装置１４は、例えばＣＲＴや液晶ディスプレイ等が使用される。この表示装置には、要約対象文書の内容や、本実施の形態により自動生成された要約の内容等が表示されるようになっている。印刷装置１５は、表示装置１４に表示された文章や、記憶装置１６の文書格納部１６４に格納された文書等の印刷を行うためのものである。この印刷装置としては、レーザプリンタ、ドットプリンタ、インクジェットプリンタ、ページプリンタ、感熱式プリンタ、熱転写式プリンタ、等の各種印刷装置が使用される。

【００１９】記憶装置１６は、読み書き可能な記憶媒体と、その記憶媒体に対してプログラムやデータ等の各種情報を読み書きするための駆動装置で構成されている。この記憶装置１６に使用される記憶媒体としては、主としてハードディスクが使用されるが、後述の記憶媒体駆動装置１７で使用される各種記憶媒体のうちの読み書き可能な記憶媒体を使用するようにしてもよい。記憶装置１６は、仮名漢字変換辞書１６１、プログラム格納部１６２、データ格納部１６３、文書データベース１６４、要約データベース１６５、文書ベクトルデータベース１６６、図示しないその他の格納部（例えば、この記憶装置１６内に格納されているプログラムやデータ等をバックアップするための格納部）等を有している。プログラム格納部１６２には、本実施の形態における自動要約処理プログラム、文書ベクトル作成処理プログラム、要約作成処理プログラム等の各種プログラムの他、仮名漢字変換辞書１６１を使用して入力された仮名文字列を漢字混り文に変換する仮名漢字変換プログラム等の各種プログラムが格納されている。データ格納部１６３には、要約パラメータのデフォルト値等の各種データが格納され

ている。要約パラメータのデフォルト値としては、例えば、全文書に対する要約の比率＝「２５％」や、日付時刻、価格情報、物理量（サイズ、重量、温度等）等の数量重視＝「しない」や、ＵＲＬ（Uniform Resource Locator）重視＝「しない」や、です／ます／であるの選択＝「しない」、等の値が格納されている。

【００２０】文書データベース１６４には、仮名漢字変換プログラムにより作成された文書や、他の装置で作成されて記憶媒体駆動装置１７や通信制御装置１８から読み込まれた文書が格納される。この文書データベース１６４に格納される各文書の形式は特に限定されるものではなく、テキスト形式の文書、ＨＴＭＬ（Hyper TextMarkup Language）形式の文書、ＪＩＳ形式の文書等の各種形式の文書の格納が可能である。文書データベース１６４には、これらの形式の文書データのが格納される。要約データベース１６５、及び文書ベクトルデータベース１６６には、文書データベース１６４に格納されている各文書に対応する要約や文書ベクトルが格納されるようになっている。

【００２１】図２は、文書ベクトルデータベース１６６の内容を概念的に表したものである。この図２に示されるように、文書中から自動抽出されたキーワードｘに対して求められた要素値ｆ（ｘ）が文書ベクトルの要素として格納されている。この文書ベクトルは各文書（Ａ、Ｂ、Ｃ…）毎に格納され、文書データベース１６４に格納されている各文書と対応づけられている。各文書ベクトルの次元は採用するキーワードｘ（重要語句）の数であるが、２文書間の類似度を両文書ベクトルから求める場合には、両文書のキーワードの和集合の数が両文書ベクトルの次元となる。この場合、一方の文書ベクトルにのみ含まれるキーワードに対する他方の文書ベクトルの要素値は、”０”に定義される。

【００２２】例えば、図２おいて、文書Ｂのキーワードは「重要、重要語、重要度、…」、文書Ｃのキーワードは「重要、…、政治、…」であり、両文書の文書ベクトルは次の通りである。

文書Ｂの文書ベクトル＝（　１，１８，１９，…）

文書Ｃの文書ベクトル＝（１８，…，２１，…）

これに対して文書Ｂと文書Ｃとの類似度を算出する場合には、両文書のキーワードを「重要、重要語、重要度、…、政治、…」とし、両文書の文書ベクトルはつぎの通り定義される。

文書Ａの文書ベクトル＝（　１，１８，１９，…，０，…）、

文書Ｃの文書ベクトル＝（１８，　０，　０，…，２１，…）

【００２３】記憶媒体駆動装置１７は、ＣＰＵ１１１が外部の記憶媒体からコンピュータプログラムや文書を含むデータ等を読み込むための駆動装置である。記憶媒体に記憶されているコンピュータプログラムには、本実施の形態の文書処理装置により実行される各種処理のためのプログラム、および、そこで使用される辞書、データ等も含まれる。ここで、記憶媒体とは、コンピュータプログラムやデータ等が記憶される記憶媒体をいい、具体的には、フロッピーディスク、ハードディスク、磁気テープ等の磁気記憶媒体、メモリチップやＩＣカード等の半導体記憶媒体、ＣＤ－ＲＯＭやＭＯ、ＰＤ（相変化書換型光ディスク）等の光学的に情報が読み取られる記憶媒体、紙カードや紙テープ等の用紙（および、用紙に相当する機能を持った媒体）を用いた記憶媒体、その他各種方法でコンピュータプログラム等が記憶される記憶媒体が含まれる。本実施の形態の文書処理装置において使用される記憶媒体としては、主として、ＣＤ－ＲＯＭやフロッピーディスクが使用される。記憶媒体駆動装置１７は、これらの各種記憶媒体からコンピュータプログラムを読み込む他に、フロッピーディスクのような書き込み可能な記憶媒体に対してＲＡＭ１１３や記憶装置１６に格納されているデータ等を書き込むことが可能である。

【００２４】本実施の形態の文書処理装置では、制御部１１のＣＰＵ１１１が、記憶媒体駆動装置１７にセットされた外部の記憶媒体からコンピュータプログラムを読み込んで、記憶装置１６の各部に格納する。そして、本実施の形態による自動要約処理等の各種処理を実行する場合、記憶装置１６から該当プログラムをＲＡＭ１１３に読み込み、実行するようになっている。但し、記憶装置１６からではなく、記憶媒体駆動装置１７により外部の記憶媒体から直接ＲＡＭ１１３に読み込んで実行することも可能である。また、文書処理装置によっては、本実施の形態の自動要約処理プログラム等を予めＲＯＭ１１２に記憶しておき、これをＣＰＵ１１１が実行するようにしてもよい。

【００２５】通信制御装置１８は、他のパーソナルコンピュータやワードプロセッサ等との間でテキスト形式やＨＴＭＬ形式等の各種形式の文書やビットマップデータ等の各種データの送受信を行うことができるようになっている。入出力Ｉ／Ｆ１９は、音声や音楽等の出力を行うスピーカ等の各種機器を接続するためのインターフェースである。文字認識装置２０は、用紙等に記載された文字をテキスト形式やＨＴＭＬ等の各種形式で認識する装置であり、イメイージスキャナや文字認識プログラム等で構成されている。

【００２６】本実施の形態では、キーボード１２の入力操作により作成した文書（ＲＡＭ１１３の所定格納エリアに格納）の他、外部で作成して所定の記憶媒体に格納した文書で記憶媒体駆動装置１７から読み込んだ文書、予め文書データベースに格納されている文書、通信制御装置１８からダウンロードした文書、及び文字認識装置２０で文字認識した文書、等の各種文書を対象文書として取得する（文字取得手段）ことが可能である。

【００２７】以上のように構成された本実施の形態の文書処理装置による、複数文書から要約を作成する自動要約処理の動作について図３から図８を用いて説明する。図３は自動要約処理のメイン動作を表したものである。図４中に示した文書ベクトルは、概念的に理解しやすくするために２次元で表示したものであるが、実際にはＮ次元ベクトルである。ＣＰＵ１１１は、要約を作成する対象となっている複数の文書の１つである要約対象文書Ａ（図４（Ａ））を取得し、ＲＡＭ１１３の要約対象文書格納エリア１１３１に格納する（ステップ１０）。要約対象文書は、ユーザの指示に従ってＲＡＭ１１３（自装置内で作成された文書である場合）、記憶装置１６の文書データベース１６４（要約が未だ作成されていない文書である場合）、記憶媒体駆動装置１７（自装置または他装置で作成済みの文書の場合）、通信制御装置１８（パソコン通信、インターネット等の通信による場合）から取得する。

【００２８】次に、ＣＰＵ１１１は、ユーザによってキーボード１２等から要約パラメータが入力された場合には入力値を取得し、ユーザによる入力がない場合にはデータ格納部１６３に格納された要約パラメータのデフォルト値を取得し、要約パラメータ格納エリア１１３２に格納する（ステップ１１）。

【００２９】次に、取得した各文書間における明示的引用関係を抽出する（ステップ１２）。ここで、明示的引用関係とは、例えば「前記…」、「…という見解がある」、「…著」など文書の外見から明白に他の文書を引用していることが明らかな引用をいう。この抽出は、特約語、特約表現知識ベースを用いて形態素解析にマッチングをかけて行う。続いて、取得した各文書間における非明示的引用関係を抽出する（ステップ１３）。非明示的引用関係とは、テキスト内容で同一のイベントに言及しつつ「…という見解もあるようだが、」のように「…」で纏められたテキストの中から、より古い記事側との類似性を判定して得られる。この文書間の類似性を判定するために、ＣＰＵ１１１は、要約対象文書格納エリア１１３１に格納した要約対象文書の各文章に対する文書ベクトルＶ（図４）を求める。

【００３０】図７は、文書ベクトル作成処理の動作を表したフローチャートである。ＣＰＵ１１１は、形態素解析を行うことで要約対象文書の文章から自立語を抽出する（ステップ１３１）と共に、名詞句、複合名詞句等を含めた候補語（句）を要約対象文書Ａから抽出しＲＡＭ１１３の所定作業領域に格納する（ステップ１３２）。そして抽出した候補語（句）の要約対象文書での出現頻度、評価関数から、各候補語（句）重要度ｆ（ｘ）を決定する（ステップ１３３）。ここで、評価関数としては、例えば、所定の重要語が予め指定されている場合にはその重要語に対する重み付け、単語、名詞句、複合名詞句等の候補語（句）の種類による重み付け等が使用される。さらにＣＰＵ１１１は、決定した重要度ｆ（ｘ）の値から要約対象文書Ａのキーワードａ，ｂ，…を決定する（ステップ１３４）。そして、各キーワードの重要度ｆ（ｘ）を要素として、文書ベクトルＶ＝（ｆ（ａ），ｆ（ｂ），…）をＲＡＭ１１３の文書ベクトル格納エリア１１３４に格納する（ステップ１３５）。この文書ベクトルＶを求める処理を複数の各文書Ｂ、Ｃ、Ｄ……と全ての要約対象文書について行う。

【００３１】要約対象の全ての文書に対して文書ベクトルＶが求まるとＣＰＵ１１１は、各文書間の類似度を求める（ステップ１３）。各文書間の類似度ｓを、両者の文書ベクトルｂｎと文書ベクトルｂｎ＋１間の角度に依存するコサインにより求める。すなわち、両文書ベクトルｂｎとｂｎ＋１間の角度をｑとし、両文書ベクトルの内積をｂｎ・ｂｎ＋１とし、両文書ベクトルの大きさをそれぞれ｜ｂｎ｜、｜ｂｎ＋１｜とした場合、両文書ベクトルの類似度ｓは次の数式１により求まる。

【００３２】

【数１】

$$\text{類似度 } s = \cos(q) = (bn \cdot bn+1) / (|bn| \times |bn+1|)$$

【００３３】この類似度ｓの値は－１≦ｓ≦１までの値をとり、１に近いほど２つの文書ベクトルが互いに平行に近く、２つの文書同士は似ていると考えることができる。図５に示すように、ＡとＢとのようにベクトルの向きが近似しているものは、内容が近似しており、Ａ、ＢとＣとのように、ベクトルの向きが異なるものは、内容が相違していることとなる。その後、ステップ１２、ステップ１３で抽出した引用関係の記述の近傍にある筆者の判断のモダリティ素性から「肯定的引用」「中立的引用」「否定的引用」などの弁別を行う（ステップ１４）。筆者の判断のモダリティ素性とは、例えば、「遺憾ながら…間違っている」「賛成できない」「反対せざるを得ない」などの記述を基に決定する。ステップ１４の弁別から各々接続詞と、トピック文、文末表現からなる「接続文」を生成する（ステップ１５）。例えば、「同様に、鈴木氏が文献Ａでも『○○○』と記述している。」「しかし、一方では、田中氏による『…文献Ａ』という意見もある。」「また、…（木村三郎氏：文献Ｂ）との報告がある。」の各文の「同様に」「しかし」「また」といった接続詞、「と記述している。」「という意見もある。」「との報告がある。」との「接続文」を用意する。

【００３４】そして、各文書の要約文を作成する（ステップ１６）。図８は、要約作成処理の動作を表したフローチャートである。ＣＰＵ１１１は、まず形態素解析を行うことで各文書群に含まれる自立語を抽出する（ステ

ップ２２１）と共に、名詞句、複合名詞句等を含めた候補語（句）を要約対象文書Ａから抽出しＲＡＭ１１３の所定作業領域に格納する（ステップ２２２）。そして、ＲＡＭ１６の要約パラメータ格納エリア１１３２に格納した要約パラメータや、抽出した候補語（句）の各文書群中での出現頻度、評価関数等から、各候補語（句）重要度ｆ（ｙ）を決定する（ステップ２２３）。ここで、評価関数としては、例えば、所定の重要語が予め指定されている場合にはその重要語に対する重み付け、単語、名詞句、複合名詞句等の候補語（句）の種類による重み付け等が使用される。

【００３５】さらにＣＰＵ１１１は、決定した重要度ｆ（ｙ）や要約パラメータ格納エリアリレーに格納された要約パラメータ等から、各文書群含まれる各センテンスに対する重要度Ｆ（ｚ）を決定する（ステップ２２４）。そして、決定したセンテンスの重要度Ｆ（ｚ）の重要度が高いセンテンスの上位から要約パラメータの要約比率（例えば、文書群の全センテンス数の内の上位２５%）以内に入るセンテンスをリストアップする（ステップ２２５）。そしてＣＰＵ１１１は、リストアップしたセンテンスを文書群の中での出現順に並べることで当該文書群についての要約とし、これをＲＡＭ１１３の要約格納エリアに格納して（ステップ２２６）、図３の自動要約処理ルーチンにリターンする。

【００３６】各文書に対する要約の作成が終了するとＣＰＵ１１１は、図６に示すように、作成された要約文の間にステップ１５で生成した接続詞、接続文を挿入して、要約文を完成する（ステップ１６）。これを要約格納エリア１１３５の所定エリアに格納して、本実施の形態による自動要約処理を終了する。以上説明したように、本実施の形態による自動要約処理によれば、複数の文書にまたがり、且つそれらの間の関係についての短い説明文を含む、可読性の高い要約を自動生成することができる。

【００３７】以上の自動要約処理が終了すると、ＣＰＵ１１１はユーザの指示によりＲＡＭ１１３に格納した各データの保存処理を行う。すなわち、要約対象文書格納エリア１１３１から要約対象文書を読み出して、記憶装置１６の文書データベース１６４に格納する。また作成した要約を要約格納エリア１１３５から読み出し、文書データベース１６４に格納した要約対象文書との関連性を付けて記憶装置１６の要約データベース１６５に格納する。さらに、文書ベクトル作成処理で求めた文書ベクトルＶを文書ベクトル格納エリア１１３５から読み出し、文書データベース１６４に格納した要約対象文書との関連性を付けて記憶装置１６の文書ベクトルデータベース１６６に格納する。

【００３８】以上、本実施の形態の構成および自動要約処理について説明したが、本発明では、これらの各形態に限定されるものではなく、請求項に記載された発明の範囲内で種々の変形をすることが可能である。例えば実施の形態では、形態素解析及び候補語（句）の抽出について、文書ベクトル作成処理（図７のステップ１３１とステップ１３２）と、要約作成処理（図８のステップ２２１とステップ２２２）とにおいて独立して同様な処理を行うこととしたが、本発明では、文書ベクトル作成処理で抽出した候補語（句）をＲＡＭ１６の所定エリアに格納しておき、要約作成処理で利用するようにしてもよい。

【００３９】また説明した実施の形態では、自動要約処理が終了した後の保存処理において、要約対象文書、要約、文書ベクトルＶのみを記憶装置１６の各データベース１６４、１６５、１６６に格納し保存するようにしたが、本発明では更に、文書ベクトル作成処理（図７）のステップ１３２で要約対象文書から抽出し、ＲＡＭ１１３の所定作業領域に格納した候補語（句）を要約対象文書Ａと関連つけて、文書データベース１６４、又は専用の候補語（句）データベースに格納するようにしてもよい。また要約パラメータ格納エリア１１３２から要約パラメータを読み出して、当該要約に関連付けて、要約データベース１６６、または専用の要約パラメータデータベースに格納するようにしてもよい。

【００４０】さらに、説明した実施の形態では、文書ベクトル作成処理（ステップ１３、図７）及び要約作成処理（ステップ２２、図８）の両処理において、形態素解析（ステップ１３１、２２１）と候補語（句）の抽出（ステップ１３２、２２２）を行った。しかし、同一センテンスに対する処理であるため、抽出した候補語（句）は同一である。そこで、本発明では、文書ベクトル作成処理で抽出した候補語（句）をＲＡＭ１１３の所定エリアに格納しておき、要約処理において格納した候補語（句）を使用することでステップ２２１とステップ２２２を省略するようにしてもよい。この候補語（句）についても、要約対象文書に対する候補語（句）として文書データベース１６４、又は専用の候補語（句）データベースに格納するようにしてもよい。

【００４１】説明した実施の形態では文書ベクトルを作成する方法として図７のフローチャートに従った方法を１例にして説明したが、本発明でこの方法に限られるものではなく、要約対象文書中Ａからキーワードを抽出する方法や、抽出キーワードに対する重要度（＝文書ベクトルの要素値）の決定方法等については、公知の各種方法により置き換えることが可能である。また、各サブ文書群に対する要約の作成処理についても同様に図８のフローチャートに示した方法に限られるものではなく、公知の各種要約方法、抄録作成方法等を資料することが可能である。更に、２つの文書ベクトルの類似度の算出方法については、数式１により類似度を算出することとしたが、この数式に限定されるものではなく、ベクトル相互間の類似関係を表すことが可能であれば他の数式によ

り類似度を算出することも可能である。

【００４２】説明した実施の形態では、日本語で作成された文書に限られるものでなく、あらゆる言語で作成された文書を対象とすることが可能である。その場合、対象となる文書が作成された言語用の形態素解析アルゴリズム等を使用するといった、本発明の構成には影響のない部分を変更するだけでよい。なお、以上の実施の形態において説明した、各装置、各部、各動作、各処理等に対しては、それらを含む上位概念としての各手段（～手段）により、実施の形態を構成することが可能である。例えば、「決定した重要度ｆ（ｘ）の値から要約対象文書Ａのキーワードａ，ｂ，…を決定する（ステップ１３４）」との記載に対して「キーワード決定手段」を構成し、「決定したセンテンスの重要度Ｆ（ｚ）の重要度が高いセンテンスの上位から要約パラメータの要約比率（例えば、サブ文書群の全センテンス数の内の上位２５％）以内に入るセンテンスをリストアップする（ステップ２２５）」との記載に対して「センテンスリストアップ手段」を構成するようにしてもよい。同様に、その他各種動作に対して「～（動作）手段」等の上位概念で実施の形態を構成するようにしてもよい。

【００４３】

【発明の効果】本発明によれば、複数の文書で構成された所定形式の文書を取得し、取得した文書をの関係を引用関係から把握して、適切な接続文を生成し、これを要約した各要約文の間に挿入することで、前後の文脈が明確な分かり易い複数文書の要約を得ることが出来る。

【図面の簡単な説明】

【図１】本発明の１実施の形態における文書処理装置の構成を表したブロック図である。

【図２】同上、実施の形態における文書ベクトルデータベースの内容を概念的に表した説明図である。

【図３】同上、実施の形態における自動要約処理のメイン動作を表したフローチャートである。

【図４】同上、実施の形態における、文書Ａに対する文書ベクトルを求めたところ示す図である。

【図５】同上、実施の形態における、各文書に対する文書ベクトルを求めたところ示す図である。

【図６】同上、実施の形態における、要約文と接続文の接合を示す図である。

【図７】同上、実施の形態における文書ベクトル作成処理の動作を表したフローチャートである。

【図８】同上、実施の形態における要約作成処理の動作を表したフローチャートである。

【符号の説明】

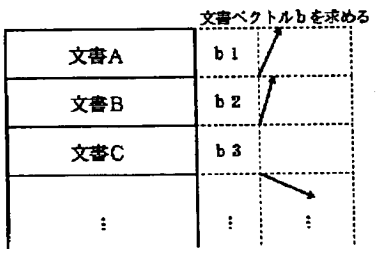
１１　制御部
１１２　ＲＯＭ
１１３　ＲＡＭ
１１３１　要約対象文書格納エリア
１１３２　要約パラメータ格納エリア
１１３３　接続文格納エリア
１１３４　文書ベクトル格納エリア
１１３５　要約格納エリア
１２　キーボード
１３　マウス
１４　表示装置
１５　印刷装置
１６　記憶装置
１６１　仮名漢字変換辞書
１６２　プログラム格納部
１６３　データ格納部
１６４　文書データベース
１６５　要約データベース
１６６　文書ベクトルデータベース
１７　記憶媒体駆動装置
１８　通信制御装置
１９　入出力Ｉ／Ｆ
２０　文字認識装置

【図２】

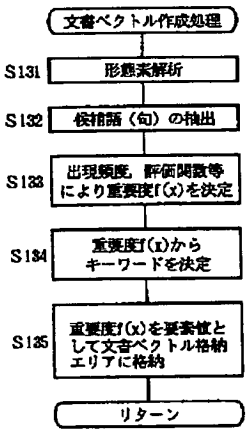
文書ベクトルデータベース

文書	キーワードの要素値ｆ（ｘ）						
	……	重要	重要語	重要度	……	政治	……
A	……	2	20	21	……	—	……
B	……	1	18	19	……	—	……
C	……	18	—	—	……	21	……
⋮							

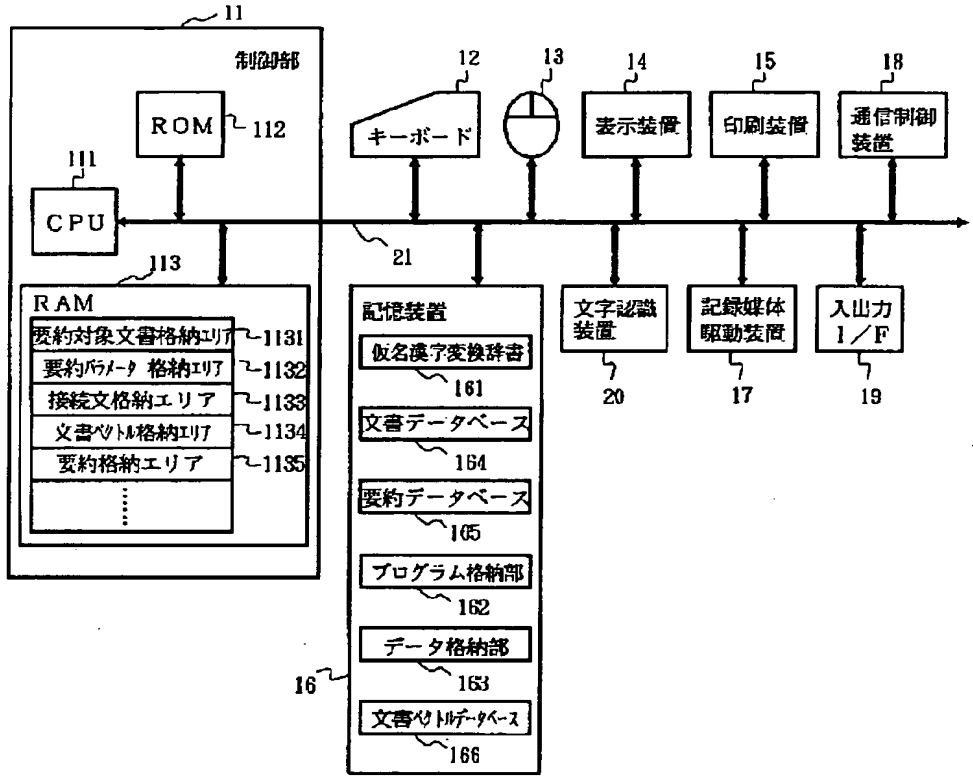
【図５】



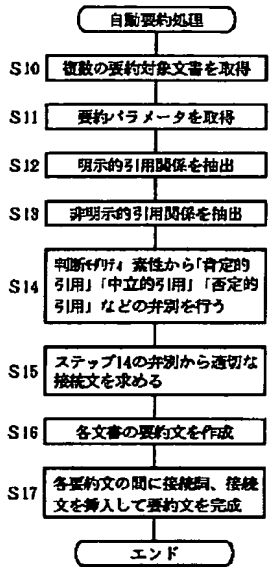
【図７】



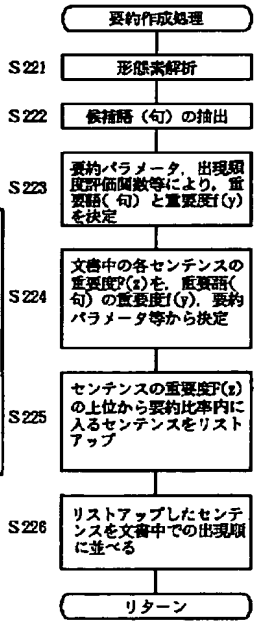
【図１】


11
制御部
ROM
112
111
CPU
113
RAM
要約対象文書格納エリア
1131
要約パラメータ 格納エリア
1132
接続文格納エリア
1133
文書ベクトル格納エリア
1134
要約格納エリア
1135
12
キーボード
13
14
表示装置
15
印刷装置
18
通信制御装置
21
記憶装置
仮名漢字変換辞書
161
文書データベース
164
要約データベース
165
プログラム格納部
162
データ格納部
163
文書ベクトルデータベース
166
16
文字認識装置
20
記録媒体駆動装置
17
入出力 I／F
19

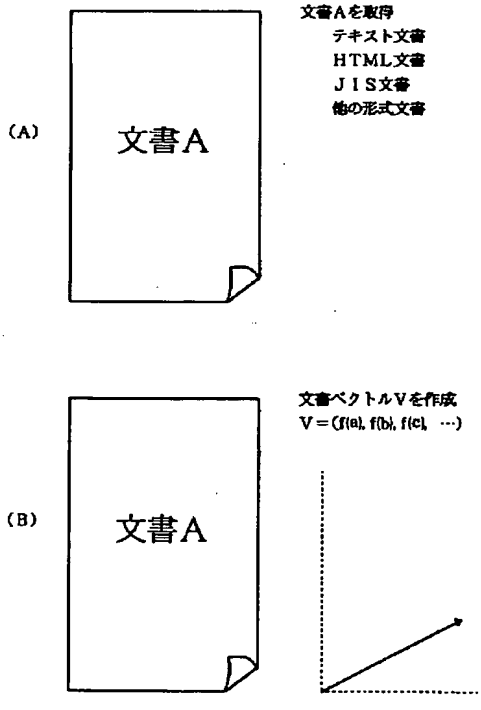
【図３】


自動要約処理
S10
複数の要約対象文書を取得
S11
要約パラメータを取得
S12
明示的引用関係を抽出
S13
非明示的引用関係を抽出
S14
判断モダリティ 表性から「肯定的引用」「中立的引用」「否定的引用」などの弁別を行う
S15
ステップ14の弁別から適切な接続文を求める
S16
各文書の要約文を作成
S17
各要約文の間に接続詞、接続文を挿入して要約文を完成
エンド

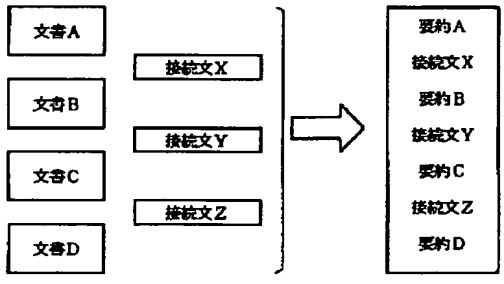
【図８】


要約作成処理
S221
形態素解析
S222
候補語（句）の抽出
S223
要約パラメータ，出現頻度評価関数等により，重要語（句）と重要度f(y)を決定
S224
文書中の各センテンスの重要度P(s)を，重要語（句）の重要度f(y)，要約パラメータ等から決定
S225
センテンスの重要度P(s)の上位から要約比率内に入るセンテンスをリストアップ
S226
リストアップしたセンテンスを文書中での出現順に並べる
リターン

【図４】


(A)
文書A
文書Aを取得
テキスト文書
HTML文書
ＪＩＳ文書
他の形式文書
(B)
文書A
文書ベクトルVを作成
V＝(f(a), f(b), f(c), …)

【図６】


文書A
文書B
文書C
文書D
接続文X
接続文Y
接続文Z
要約A
接続文X
要約B
接続文Y
要約C
接続文Z
要約D